Part 1, Section 4.1: Understanding Risk and Adverse Impacts

The definition of risk in subsection 4.1 is inconsistent. First, the paper introduces risk as a “measure of the extent to which an entity is negatively influenced by a potential circumstance or event.” The following paragraph cites IEC/ISO 31010 to say that certain risks can be positive. These two statements are incompatible, though neither is necessarily wrong (that choice or debate is left to the authors of this initial draft). In this subsection, the former definition of risk follows from the notion that risk can be viewed as an event with a negative impact, in contrast with the definition of opportunity to mean any event that can have a positive impact (from [2]). The latter definition is compatible with the cited IEC/ISO 31010, which defines risk to include the effects of types of uncertainty that may lead to positive or negative consequences. Perhaps this draft could use a subsection on normative references that point to documents for which terms such as risk, risk management, artificial intelligence, etc., can be defined and used consistently.

Furthermore, the draft asserts that risk is a function of two components: the adverse impacts that arise if an event occurs and the likelihood of occurrence. Another component of the risk function is the degree of knowledge available about the event that results in adverse impacts. For example, knowledge of the event may include underlying variables or factors for which the adverse impacts from the event is conditional on. The availability and extent of event-related knowledge affects the risk assessment. Risk should be extended and specified as also a function of the body of knowledge used to assess risk [1].

Part 1, Section 5: AI Risks & Trustworthiness

A central premise of this draft is that increasing the trustworthiness of AI tools can facilitate their adoption and effective use. This paper introduces the following characteristics of an AI tool that supports its trustworthiness: accuracy, explainability, interpretability, privacy, reliability, robustness, safety, and mitigation of harmful bias. In addition, fairness, accountability, and transparency serve as guiding principles for trustworthiness.

Several issues stand out with Section 5’s handling of these characteristics and guiding principles.

First, the three-tier taxonomy seems to link the technical and socio-technical characteristics with the guiding principles. The paper presents the taxonomy as a framework for which each of its comprising objects (from accuracy to transparency) has associated risks. For example, an AI model has an accuracy characteristic, and there are risks associated with the model’s accuracy. Do the guiding principles also serve as characteristics of AI tools for which there are risks associated? This paper should make this distinction more explicitly or consider an alternate framing. For example, an alternative approach could show the effect that alleviating each risk associated with each (technical or socio-technical) characteristic has on fairness, accountability, and traceability. As a side note, the paper should also consider mentioning that the distinction between technical and socio-technical characteristics may not be as clear-cut as metrics for some of the designated socio-technical issues are being developed and automated [3].

Another issue in this section regards the definitions for the AI tool characteristics that affect their trustworthiness. Bringing in other standards and cited literature will help keep consistent terminology throughout the AI RMF. This is especially true for terms that are often confounded, such as explainability and interpretability. In this case, definitions for these two conflict with key literature, where explainability is the ability of an approximate model to explain an original closed-box model, and interpretability is the degree to which a model or its data obeys domain-specific constraints to be more easily understood [5]. Citations can convey the draft’s intentions with terminology choice.

Lastly, this is part of a broader discussion of how Part 1 of this draft can establish the context for AI risk management. It may be of great use to consider the most prevalent risks associated with each characteristic. Each characteristic’s subsection mentions examples of

risks. Still, a more comprehensive list of the significant risks may challenge the community to address them (as was done by focusing on lists of major CBRN threats in the security community). The challenge here is that risks associated with a particular AI tool characteristic don't necessarily transfer between different domains of AI tool application. Another aside: mentions of data drift and concept drift issues can be presented as risks in subsections 5.1.2 and 5.1.3 [4].

Part 2 & Part 3

The stated purpose of Part 2 is to guide activities to carry out the AI risk management process. The AI RMF Core presented in this section has many overlaps with existing risk management framework standards without any reference to them. For example, the risk management processes outlined in ISO 31000 and ISO 31010 include risk treatment, risk assessment, recording and reporting, monitoring and review, communications and consultations, and scope, context, and criteria. These processes can be traced back to the AI RMF Core activities. This may facilitate understanding of the AI RMF activities for practitioners who are already familiar with existing standards and terminology for risk management frameworks and processes. And it reflects Section 1 of the AI RMF, which states that there are many commonalities between managing risks for AI and risks of other types of technologies.

Overall, the authors may consider strengthening Part 2 by describing how the risks of AI-based tools differ from that of other technologies. This section can explain how these differences impact the existing risk management frameworks and processes (for example, from ISO 31000), resulting in the AI RMF Core. It would also be interesting to see how risk assessment techniques (appendices A & B of ISO 31010, used during ISO 31000 risk management processes) adapt to AI-based tool risks.

Lastly, as Part 3 is meant to offer sample practices to be considered in carrying out some of the guidance in the AI RMF, we suggest taking a peek at our work in [6]. This work steps through a risk-based evaluation method to understand the impact of AI-based tools used for monitoring manufacturing processes.

[1] Zio, Enrico. "The future of risk assessment." *Reliability Engineering & System Safety* 177 (2018): 176-190.
[2] PricewaterhouseCoopers, L. L. P. "Committee of Sponsoring Organizations of the Treadway Commission [COSO].(2004)." *Enterprise risk management: Integrated framework*.
[3] Rudin, Cynthia, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. "Interpretable machine learning: Fundamental principles and 10 grand challenges." *Statistics Surveys* 16 (2022): 1-85.
[4] Ng, Andrew. "Issue 102: Face Recognition Audit, Gamers Cheat with AI, Who Rules the Smart City?" The Batch. July 28, 2021. https://read.deeplearning.ai/the-batch/issue-102/.
[5] Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1, no. 5 (2019): 206-215.
[6] Sharp, Michael, Mehdi Dadfarnia, Timothy Sprock, and Douglas Thomas. "Procedural Guide for System-Level Impact Evaluation of Industrial Artificial Intelligence-Driven Technologies: Application to Risk-Based Investment Analysis for Condition Monitoring Systems in Manufacturing." *Journal of Manufacturing Science and Engineering* 144, no. 7 (2022).